QUANTILE-DOMAIN MULTIPLE INSTANCE LEARNING FOR WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide images (WSIs) contain billions of pixels, making end-to-end modeling costly in computation and storage. Their classification is therefore typically formulated as Multiple Instance Learning (MIL) over patch features. However, some methods construct bag-level context with positional structures, neighborhoods, or approximations that depend on patch indices. Their predictions can therefore change with patch order. We show that patch interactions need not cause this sensitivity and that it can instead arise from index-addressed operators. Based on this insight, we introduce Quantile-Domain Multiple Instance Learning (QDMIL), a three-stage framework. First, it forms a fixed-size, permutation-invariant quantile grid over multiple learnable projections. This grid captures the slide-wide distribution of morphological features. Second, QDMIL uses the grid to derive soft quantile coordinates for each patch and bag-shared affine modulation. Together, these signals contextualize the patch representations. Finally, QDMIL fuses attention-pooled local discriminative evidence with a separate grid encoding for slide classification. We establish sufficient operator-level conditions for permutation invariance and prove that QDMIL satisfies them under deterministic inference and exact arithmetic. Against 15 baselines across four datasets, QDMIL achieves the highest mean accuracy and F1-Score in all eight dataset–metric comparisons. Paired order-control experiments across seven methods further support distribution-relative patch positioning as a general principle for permutation-invariant bag-level context.

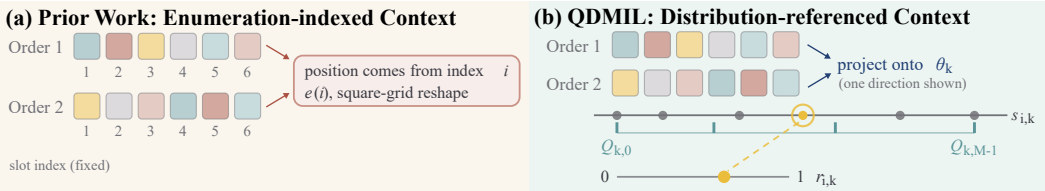

Figure 1: **Enumeration-indexed versus distribution-referenced context.** Colors denote the same patches under two input orders. **(a)** Context tied to enumeration positions. **(b)** Context defined relative to a bag-level quantile grid.

1 INTRODUCTION

Whole-slide images (WSIs) are widely used for disease diagnosis, classification, and grading in computational pathology (Bulten et al., 2022; Brancati et al., 2022). However, their gigapixel scale makes end-to-end modeling costly, while patch- or pixel-level annotations require substantial expert effort (Campanella et al., 2019). Multiple Instance Learning (MIL) instead casts a WSI as a bag of patches and can be trained using only the slide label (Ilse et al., 2018; Li et al., 2021). A common implementation embeds the patches with a frozen pathology foundation model and pools their embeddings to predict the slide label (Lu et al., 2024; Shao et al., 2026). This formulation reduces reliance on whole-slide pixel-level modeling and dense supervision, but within-slide heterogeneity remains a central challenge. Specifically, diagnostic evidence may be localized to a few regions or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distributed across multiple morphological patterns (Couture et al., 2018). Consequently, MIL methods must identify discriminative patches while using bag-level context to assess their relevance to the slide as a whole.

Existing MIL methods introduce bag-level context in several ways. Attention- and reference-based methods score patches independently or relative to bag-selected instances before aggregation (Ilse et al., 2018; Li et al., 2021; Chikontwe et al., 2025). Interaction-based methods use self-attention or learned global representations to exchange information across the bag (Shao et al., 2021; Zhu et al., 2024). Structure-based methods organize patches through spatial layouts, graphs, or hypergraphs to capture local and higher-order relationships (Fang et al., 2024; Chen et al., 2025). Despite these different designs, some implementations tie positional structures, neighborhoods, or approximate computations to the input sequence, making their predictions sensitive to patch order.

To understand this sensitivity, we examine how context is assigned to individual patches. Figure 1(a) makes the issue visible: the bag contains the same patches after reordering, yet enumeration-indexed operators assign them a different contextual structure. This observation separates two questions that are often conflated: whether patches interact and how they are addressed during interaction. Patch interactions are not inherently order dependent. Order sensitivity can arise when context is assigned through enumeration indices, which change when the input sequence is permuted. In contrast, physical coordinates remain attached to their patches, and permutation-invariant bag statistics remain unchanged. Figure 1(b) illustrates the resulting motivation: using the within-bag distribution as an order-independent reference for contextualizing patches.

We therefore propose Quantile-Domain Multiple Instance Learning (QDMIL). While preserving permutation invariance, QDMIL jointly models local discriminative evidence and whole-slide morphological composition. To this end, QDMIL is built around a single bag-level quantile grid. First, it projects the patch features along multiple learnable directions and compresses the resulting score distributions into a fixed-size bag-level quantile grid. The grid summarizes slide-wide morphological variation across the learned directions. Second, it uses the same grid to compute soft quantile coordinates and generate bag-shared affine modulation for patch updates. The updated features thus incorporate each patch's position within the bag distribution and the slide's overall distributional state. Finally, it fuses an attention-pooled representation of local evidence with a separately encoded representation of the quantile grid for slide classification. The shared grid thus couples patch-level contextualization with slide-level distribution modeling.

This work reframes permutation invariance in MIL as an addressing problem rather than a restriction on contextual interactions. We formalize this view through sufficient operator-level conditions for permutation invariance. We then prove the invariance of the quantile grid and bag-level prediction, together with the equivariance of patch updates. Across four datasets, QDMIL achieves the highest mean ACC and F1-Score in all eight dataset–metric comparisons against 15 baselines. Controlled reordering experiments further support the empirical order robustness of contextual interactions when patches are addressed through order-independent references. Together, these results position the within-bag distribution as a natural interface between local evidence and global composition. This perspective provides a broader principle for order-robust set modeling in computational pathology.

2 Related work

Weakly supervised WSI modeling. Research on WSI classification with slide-level supervision has expanded from selecting diagnostic patches to representing heterogeneous evidence across an entire slide. Attention pooling established a permutation-invariant mechanism for learning the contribution of each instance (Ilse et al., 2018). Pathology-specific extensions then introduced instance clustering, critical-instance references, and hierarchical pseudo-bags (Lu et al., 2021; Li et al., 2021; Zhang et al., 2022). The field has since broadened beyond the pooling operator itself. Recent WSI methods target instance attribution and attention stability (Cai et al., 2025; Ye et al., 2026). Related MIL work also enriches bag representations with vision–language semantics (Zhang et al., 2026). A recent study further identifies task-specific patch adaptation as an overlooked bottleneck (Shao et al., 2026). Together, these developments make contextualization before pooling a central question. A model must determine not only which patches are informative, but also what each patch means relative to the rest of the slide.

Contextual and structural MIL. A major line of work enriches patch representations with information beyond independent instance scoring. Transformer and global-representation methods construct context through cross-patch interactions or learned slide-level tokens (Shao et al., 2021; Zhu et al., 2024). Spatial, graph, and hypergraph methods instead encode tissue neighborhoods and higher-order organization (Fang et al., 2024; Chen et al., 2025). Despite their architectural diversity, these methods can be understood along two distinct axes. One is the source of context, such as feature similarity, spatial topology, or a learned global representation. The other is the addressing rule that associates this context with individual patches. Prior work has primarily enriched the source of context to capture broader dependencies or tissue structure. Less attention has been paid to how that context is assigned back to individual patches. When this assignment follows sequence indices, reordering the patches can change the computation even though the bag itself is unchanged.

Distribution as an order-independent reference. For unordered inputs, Deep Sets uses symmetric aggregation, while Set Transformer allows patch interactions without relying on enumeration order (Zaheer et al., 2017; Lee et al., 2019). Ordered statistics provide a complementary route because value ranks and quantiles depend on the empirical distribution rather than enumeration order. In this direction, Couture et al. (2018) represent images using quantiles of regional predictions, while FSPool learns over feature-wise sorted positions (Zhang et al., 2020). These approaches use distributional structure primarily to form a bag representation. FR-MIL instead recalibrates patches relative to the highest-response instance, but this reference captures only one point in the bag distribution (Chikontwe et al., 2025). QDMIL bridges these two roles by constructing a shared quantile grid over multiple learned projections. The grid provides distribution-relative coordinates for patch contextualization and a separate encoding of slide-wide composition for classification. QDMIL therefore positions the within-bag distribution as both an order-independent addressing system and a global representation.

3 QDMIL

In MIL, a WSI is represented as a bag $\mathcal{B} = \{h_i\}_{i=1}^{N}$ (Ilse et al., 2018). Each $h_i \in \mathbb{R}^d$ is the feature of patch i extracted by a frozen encoder (Lu et al., 2021). Only the slide-level label $y \in \{1, \dots, C\}$ is available during training, where C denotes the number of classes. For matrix operations, the patch features are stacked according to the order provided by the data pipeline as $H = [h_1, \dots, h_N]^\top \in \mathbb{R}^{N \times d}$. QDMIL maps H to updated patch representations through $\mathcal{G} : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times h}$ and produces class logits through $F : \mathbb{R}^{N \times d} \to \mathbb{R}^C$. For any $N \times N$ permutation matrix P, we require

$$\mathcal{G}(PH) = P\mathcal{G}(H), \qquad F(PH) = F(H). \tag{1}$$

The first condition preserves patch correspondence after reordering, while the second keeps the slide prediction unchanged.

As shown in Figure 2, the shared grid gives patches a slide-specific but order-independent reference and connects patch contextualization to the global readout.

Proposition 1 (Sufficient conditions for permutation equivariance and invariance). Consider a model that applies the same mapping to every patch and forms all bag-level quantities through operations independent of enumeration order. If the same function updates each patch from its own representation and these bag-level quantities, the patch update is permutation equivariant. If the final prediction depends only on order-independent aggregation results, the bag-level prediction is permutation invariant. Under deterministic inference and exact arithmetic, $\mathcal{G}$ and F therefore satisfy Eq. (1). Summation, averaging, extrema, and quantiles are examples of such order-independent aggregation operations. The proof is provided in Appendix B.2.

Sections 3.1–3.3 instantiate these conditions through quantile-grid construction, quantile-conditioned patch updates, and bag-level representation fusion. Bias terms in learned linear layers are omitted for brevity.

3.1 Bag-level Quantile Grid

High-dimensional patch features have no natural total order, so their quantiles cannot be defined directly. Quantiles of regional predictions and feature-wise sorting nevertheless provide order-independent summaries of visual sets (Couture et al., 2018; Zhang et al., 2020). QDMIL makes

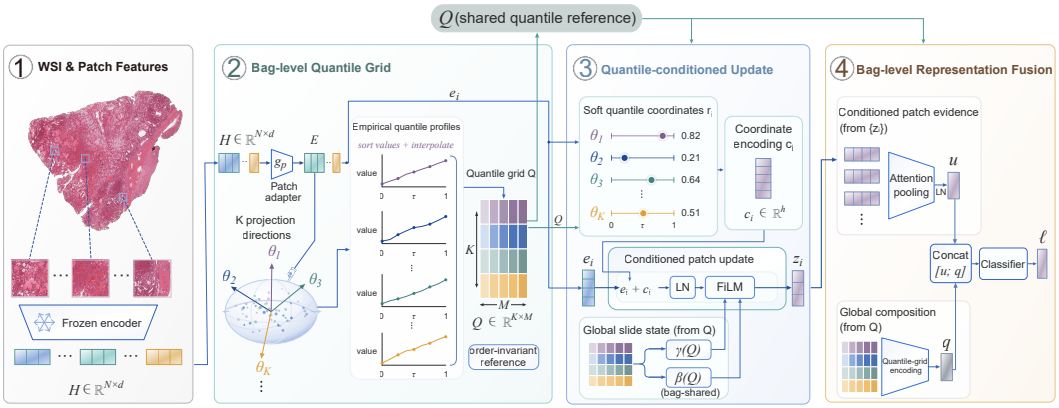

Figure 2: **Overview of QDMIL.** (1) A frozen encoder transforms WSI patches into a feature bag H. (2) A shared patch adapter and K learnable projections form the bag-level quantile grid Q. (3) The grid provides soft quantile coordinates and bag-shared modulation to update e_i into z_i. (4) Attention-pooled patch evidence u and the grid encoding q are fused for classification.

this principle applicable to patch features by learning K scalar projection views and representing each view through its empirical quantiles. For a WSI, these profiles summarize slide-wide variation in task-relevant morphological features.

We first adapt the frozen patch features to the task space using $e_i = g_p(h_i) \in \mathbb{R}^h$ and collect them as $E = [e_1, \ldots, e_N]^\top \in \mathbb{R}^{N \times h}$. Specifically, $g_p(x) = \mathrm{Dropout}(\mathrm{GELU}(Ax))$, where $A \in \mathbb{R}^{h \times d}$ is learnable. To parameterize the projection views, let $W = [W_1, \ldots, W_K] \in \mathbb{R}^{h \times K}$ have nonzero columns and normalize each column as $\theta_k = W_k / \|W_k\|_2$. This normalization prevents the magnitude of W_k from scaling the projected scores. Collecting the unit directions as $\Theta = [\theta_1, \ldots, \theta_K] \in \mathbb{R}^{h \times K}$ gives the score matrix $S = E\Theta \in \mathbb{R}^{N \times K}$, whose (i, k)-th entry is

$$s_{i,k} = \langle e_i, \theta_k \rangle. \tag{2}$$

For each k, the multiset $\{s_{i,k}\}_{i=1}^N$ defines a one-dimensional empirical distribution of the slide. The K directions therefore provide complementary views of the within-bag feature distribution. They are not assigned predefined pathological meanings, but are learned from slide-level labels to expose variation across the patch population.

We next convert each variable-size empirical distribution into one fixed-size row of the grid. Let $G_k : [0, 1] \to \mathbb{R}$ denote the linearly interpolated quantile function of the projected scores along direction k. We sample G_k at $M \geq 2$ uniformly spaced quantile levels,

$$\tau_m = \frac{m}{M-1}, \qquad Q_{k,m} = G_k(\tau_m), \quad m = 0, \ldots, M-1. \tag{3}$$

Computationally, G_k is obtained by sorting the N projected scores and linearly interpolating between adjacent order statistics. Each τ_m is mapped to the zero-based position $p_m = \tau_m(N-1)$, and $Q_{k,m}$ is obtained by interpolation at p_m. The exact form and boundary cases are given in Appendix B.1. Because $\tau_0 = 0$ and $\tau_{M-1} = 1$, the grid endpoints satisfy $Q_{k,0} = \min_i s_{i,k}$ and $Q_{k,M-1} = \max_i s_{i,k}$, preserving the extrema of each projection view.

The levels $\{\tau_m\}_{m=0}^{M-1}$ are fixed and shared across bags, while the input mapping and projection directions are learned. Stacking the sampled quantiles from all directions forms $Q \in \mathbb{R}^{K \times M}$. Its size is independent of N, and it provides a fixed-size summary of the slide-wide projected morphological feature distributions. Because sorting and interpolation depend only on the multiset of scores, Q is invariant to patch enumeration.

By sampling intermediate quantiles, the grid can retain distributional differences that are not captured by means and extrema alone. At the same time, its approximation error is controlled despite using only M values per direction. For $N \geq 2$ and $M \geq 2$, the distribution defined by the piecewise-linear quantile function reconstructed from the grid differs from the empirical distribution of projected scores by at most $R_k/(2N) + R_k/[2(M-1)]$ in 1-Wasserstein distance, where

$R_k = \max_i s_{i,k} - \min_i s_{i,k}$. The two terms account for quantile interpolation and grid discretization, respectively. The bound thus characterizes the accuracy of the fixed-size distributional reference. A worked example, formal statements, and proofs are provided in Appendix B.2.

3.2 Quantile-conditioned Update

In a WSI, the slide-level relevance of a local morphological pattern can depend on whether it is unusual or prevalent relative to the slide's patch population. QDMIL captures this context in feature space by locating each patch within the slide-specific quantile grid. For patch i and direction k, we assign weights to the quantile points based on their squared distances from $s_{i,k}$ and average the corresponding quantile levels,

$$r_{i,k} = \sum_{m=0}^{M-1} \mathrm{softmax}_m\left(-\frac{(s_{i,k} - Q_{k,m})^2}{T}\right) \tau_m \in [0,1]. \tag{4}$$

Here, $\mathrm{softmax}_m$ is normalized only along the quantile-level dimension, so each patch obtains a separate set of weights for each projection direction. The temperature T is parameterized by a learnable scalar t as $T = \min\{10, \max\{10^{-3}, \exp(t)\}\}$. We initialize $t = \log T_0$, where T_0 is the initial temperature, and use the learned temperature at inference. A smaller T concentrates the weights more strongly on nearby quantile points. We refer to $r_{i,k}$ as the soft quantile coordinate of patch i along direction k, and collect these coordinates into $r_i \in [0,1]^K$. A coordinate near 0 or 1 places the patch near the lower or upper tail of the corresponding slide-specific distribution. These coordinates describe distributional position rather than physical location in the WSI.

Together, r_i and Q drive a patch-specific coordinate encoding and bag-shared affine modulation. Flattening Q direction by direction gives $v_Q = (Q_{1,0}, \ldots, Q_{1,M-1}, \ldots, Q_{K,0}, \ldots, Q_{K,M-1})^\top \in \mathbb{R}^{KM}$. Using separate parameter sets, the coordinate encoding c_i and affine coefficients γ, β are computed and applied as

$$\begin{aligned} c_i &= B_{r,2}\,\mathrm{GELU}(B_{r,1} r_i), \\ [\gamma;\beta] &= B_{g,2}\,\mathrm{GELU}(B_{g,1} v_Q), \\ z_i = g_c(e_i, r_i, Q) &:= \mathrm{LayerNorm}(e_i + c_i) \odot (1+\gamma) + \beta. \end{aligned} \tag{5}$$

Here, h' denotes the hidden width of the coordinate and modulation networks, with $B_{r,1} \in \mathbb{R}^{h' \times K}$, $B_{r,2} \in \mathbb{R}^{h \times h'}$, $B_{g,1} \in \mathbb{R}^{h' \times KM}$, and $B_{g,2} \in \mathbb{R}^{2h \times h'}$. The concatenated vector $[\gamma;\beta] \in \mathbb{R}^{2h}$ is split into $\gamma, \beta \in \mathbb{R}^h$, and $\odot$ denotes element-wise multiplication. LayerNorm normalizes each patch representation across its feature dimension.

The update adds the patch-specific encoding c_i to the patch feature, applies LayerNorm, and then uses the bag-shared γ and β for FiLM-style feature-wise affine modulation (Perez et al., 2018). The factor $1 + \gamma$ parameterizes scaling as a residual adjustment to unit scale. The resulting z_i can therefore differ for similar local patterns when their distribution-relative positions or slide-wide morphological feature distributions differ.

3.3 Bag-level Representation Fusion

WSI classification must accommodate localized discriminative patterns and broader morphological variation across the slide (Couture et al., 2018). QDMIL uses two readouts: attention summarizes the conditioned patches, while a separate grid encoding summarizes the slide distribution without patch-level attention weighting. First, the conditioned patch representations z_i are pooled by gated attention (Ilse et al., 2018) and normalized,

$$\begin{aligned} a_i &= w^\top(\tanh(V z_i) \odot \sigma(U z_i)), \\ \alpha_i &= \frac{\exp(a_i)}{\sum_{j=1}^N \exp(a_j)}, \qquad u = \mathrm{LayerNorm}\left(\sum_{i=1}^N \alpha_i z_i\right). \end{aligned} \tag{6}$$

Here, σ denotes the sigmoid function, and a denotes the attention module's hidden dimension. The parameters satisfy $V, U \in \mathbb{R}^{a \times h}$ and $w \in \mathbb{R}^a$. The attention weights are normalized only along the patch dimension. The LayerNorm applied to the bag representation and the patch-level LayerNorm

in Eq. (5) use separate learnable affine parameters. The resulting representation u allows strongly weighted morphological patterns to influence the slide prediction even when they occur in relatively few patches.

To complement this selective summary, we encode v_Q into a separate distributional representation q. Because this branch operates directly on Q, it provides a fixed-size summary of slide-wide variation in the learned morphological features. The two representations are concatenated and linearly mapped to class logits,

$$\begin{aligned} q &= \mathrm{Dropout}(\mathrm{GELU}(B_q v_Q)), \\ \ell &= F(H) = B_o[u; q]. \end{aligned} \tag{7}$$

Here, $u, q \in \mathbb{R}^h$, $B_q \in \mathbb{R}^{h \times KM}$, and $B_o \in \mathbb{R}^{C \times 2h}$. The notation $[\,\cdot\,;\,\cdot\,]$ denotes concatenation along the feature dimension, and $\ell \in \mathbb{R}^C$ contains the class logits.

The complete model is trained using slide-level labels with the multiclass cross-entropy loss

$$\mathcal{L}(H, y) = -\log\big(\mathrm{softmax}(\ell)_y\big). \tag{8}$$

The pathology encoder remains frozen, while all trainable parameters of QDMIL are jointly optimized. The quantile levels are fixed, and sorting and linear interpolation are differentiable almost everywhere. Details of gradient propagation are provided in Appendix B.1.

QDMIL does not construct an $N \times N$ patch-relation matrix. With d, K, M, h, h', and a fixed, its forward-pass time complexity is $O(N \log N)$, with sorting determining the asymptotic order. The soft-assignment tensor has size $N \times K \times M$, and intermediate storage scales linearly with N under these fixed dimensions. Per-module costs and empirical runtime and memory measurements are provided in Appendix D.

Corollary 2 (Permutation invariance of QDMIL). Under deterministic inference and exact arithmetic, the conditioned patch sequence produced by QDMIL is permutation equivariant and the bag-level prediction is invariant to any permutation of the enumeration order. This follows because Q is invariant, the coordinate and update mappings are shared patch-wise, and both u and q are invariant under patch permutation. Their fusion therefore satisfies Proposition 1. The full proof is provided in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate QDMIL on TCGA-KIRP for binary early-versus-advanced stage classification (Cancer Genome Atlas Research Network, 2016), PANDA for six-class ISUP grading (Bulten et al., 2022), and Brain-Mets (Chadha et al., 2026) and BRACS (Brancati et al., 2022) for three-class classification. Estimated median bag sizes range from approximately 10^2 patches for PANDA to 10^3–10^4 for the other datasets. All methods share 512-dimensional features extracted by frozen CONCH (Lu et al., 2024) from patches obtained through the CLAM preprocessing pipeline (Lu et al., 2021). Dataset statistics, preprocessing details, and data splits are provided in Appendix A.

Baselines and evaluation. We compare QDMIL with 15 baselines, including ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), DSMIL (Li et al., 2021), CLAM-SB and CLAM-MB (Lu et al., 2021), DTFD-MIL (Zhang et al., 2022), RankMix (Chen & Lu, 2023), IBMIL (Lin et al., 2023), DGR-MIL (Zhu et al., 2024), ACMIL (Zhang et al., 2024), FR-MIL (Chikontwe et al., 2025), DyHG (Chen et al., 2025), GVLMIL (Zhang et al., 2026), and ASMIL-EMA and ASMIL (Ye et al., 2026). All methods are trained for a maximum of 100 epochs. Following prior computational pathology studies that use fold based evaluation (Lu et al., 2021; Shao et al., 2021; Zhang et al., 2022), we train every method from scratch on the same five folds. All methods use the same training seed of 3407 throughout the five fold evaluation. Given the computational demands of training on WSIs, this single seed protocol is common practice in computational pathology. Finally, we report ACC and F1-Score as mean$\pm$std across folds.

4.2 MAIN RESULTS

Under the shared encoder and evaluation protocol, QDMIL achieves the highest mean ACC and F1-Score in all eight dataset–metric comparisons (Table 1). It is the only method to do so, yielding an average rank of 1.00. With equal dataset weighting, QDMIL exceeds CLAM-MB, the strongest overall baseline, by 2.19 percentage points in ACC and 2.30 in F1-Score. Its improvement in both metrics indicates that the advantage is not confined to prevalent classes or a single pathological setting. Some baselines approach QDMIL on individual datasets or metrics, but their results are less consistent across the full comparison. QDMIL separates local evidence selection from distributional context. Attention retains localized diagnostic evidence, while the fixed-size quantile grid represents slide-wide morphology and locates each patch within that distribution without using patch enumeration. Appendix E.3 further examines performance across PANDA grades, folds, bag sizes, and ordinal errors. QDMIL therefore provides a stable, permutation-invariant integration of local selectivity and global pathological context across heterogeneous WSI tasks.

Table 1: Comparison on four datasets. Values are mean±std over five folds, and bold marks the highest mean. Avg. Rank averages the eight dataset–metric ranks, with lower values indicating better overall performance.

Method	TCGA-KIRP		PANDA		BRACS		Brain-Mets		Avg. Rank
	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	ACC	F1-Score	
ASMIL '26	0.8655 ±0.0574	0.8098 ±0.0781	0.7198 ±0.0075	0.6671 ±0.0162	0.8149 ±0.0646	0.7221 ±0.0692	0.9020 ±0.0705	0.8421 ±0.1131	8.31
ASMIL-EMA '26	0.8655 ±0.0508	0.8175 ±0.0676	0.7238 ±0.0032	0.6710 ±0.0045	0.8250 ±0.0545	0.7108 ±0.0270	0.9020 ±0.0839	0.8575 ±0.1074	7.38
GVLMIL '26	0.8579 ±0.0534	0.8074 ±0.0702	0.7201 ±0.0082	0.6635 ±0.0140	0.8250 ±0.0599	0.6871 ±0.1057	0.9285 ±0.0500	0.8852 ±0.0874	7.81
DyHG '25	0.8655 ±0.0574	0.8191 ±0.0721	0.7471 ±0.0055	0.7046 ±0.0070	0.8072 ±0.0398	0.7110 ±0.0403	0.9194 ±0.0359	0.8907 ±0.0380	6.19
FR-MIL '25	0.7983 ±0.0721	0.7284 ±0.0652	0.6902 ±0.0046	0.6270 ±0.0061	0.7920 ±0.0632	0.6421 ±0.0407	0.7403 ±0.1350	0.6369 ±0.1652	15.75
ACMIL '24	0.8654 ±0.0558	0.8254 ±0.0597	0.7181 ±0.0043	0.6653 ±0.0076	0.8402 ±0.0508	0.7417 ±0.0767	0.9016 ±0.0651	0.8496 ±0.1107	6.25
DGR-MIL '24	0.8693 ±0.0465	0.8253 ±0.0529	0.7103 ±0.0051	0.6512 ±0.0085	0.8250 ±0.0498	0.6594 ±0.1053	0.8830 ±0.0240	0.7935 ±0.0675	9.81
IBMIL '23	0.8280 ±0.0621	0.7617 ±0.0939	0.7155 ±0.0040	0.6603 ±0.0040	0.8225 ±0.0442	0.7387 ±0.0692	0.8925 ±0.0491	0.8432 ±0.0780	11.12
RankMix '23	0.8430 ±0.0603	0.7850 ±0.0671	0.7092 ±0.0017	0.6490 ±0.0090	0.8428 ±0.0421	0.7315 ±0.0762	0.8565 ±0.0458	0.7289 ±0.1234	11.12
DTFD-MIL '22	0.8691 ±0.0468	0.8198 ±0.0506	0.6605 ±0.0072	0.5632 ±0.0149	0.8351 ±0.0293	0.7138 ±0.0287	0.8834 ±0.0495	0.7949 ±0.1093	10.00
CLAM-MB '21	0.8692 ±0.0407	0.8257 ±0.0505	0.7252 ±0.0075	0.6740 ±0.0091	0.8326 ±0.0522	0.7537 ±0.0460	0.9194 ±0.0359	0.8807 ±0.0473	3.44
CLAM-SB '21	0.8653 ±0.0438	0.8012 ±0.0663	0.7191 ±0.0045	0.6614 ±0.0094	0.8377 ±0.0400	0.7463 ±0.0731	0.9193 ±0.0482	0.8651 ±0.0829	7.00
DSMIL '21	0.8691 ±0.0284	0.8217 ±0.0318	0.7062 ±0.0052	0.6444 ±0.0083	0.8275 ±0.0523	0.7212 ±0.0984	0.8656 ±0.0688	0.7784 ±0.1284	10.06
TransMIL '21	0.8504 ±0.0629	0.8076 ±0.0697	0.7161 ±0.0038	0.6620 ±0.0078	0.8250 ±0.0369	0.6569 ±0.0937	0.8482 ±0.0992	0.6883 ±0.2569	12.06
ABMIL '18	0.8430 ±0.0603	0.7850 ±0.0671	0.7171 ±0.0057	0.6632 ±0.0049	0.8351 ±0.0354	0.7277 ±0.0559	0.9107 ±0.0524	0.8558 ±0.0851	8.69
QDMIL (ours)	**0.8878** ±0.0389	**0.8302** ±0.0576	**0.7636** ±0.0058	**0.7244** ±0.0105	**0.8453** ±0.0346	**0.7566** ±0.0720	**0.9372** ±0.0400	**0.9147** ±0.0561	**1.00**

4.3 ABLATION STUDY

We ablate the two pathways through which QDMIL uses slide-level distributional context: the conditioned patch update and the separate quantile-grid encoding. We also remove quantile coordinates or FiLM modulation from the update, while retaining patch-level LayerNorm and the grid-encoding branch. Removing quantile coordinates sets $c_i = 0$, and removing FiLM sets $\gamma = \beta = 0$. The full model achieves the highest mean ACC and F1-Score on both datasets (Table 2), supporting com-

plementary contributions from the two pathways. The pattern differs between pathology tasks. The conditioned update contributes more on TCGA-KIRP. On PANDA, the grid encoding is more important, and its removal reduces ACC and F1-Score by 3.20 and 4.36 percentage points. This larger effect is consistent with prostate grading relying on the slide-wide composition of Gleason patterns. Within the patch update, removing FiLM causes larger reductions than removing quantile coordinates on both datasets. Bag-shared modulation therefore carries much of the distributional signal used to contextualize local morphology. The results support both roles of the quantile grid. It conditions patch evidence and provides a slide-wide representation for classification. Appendix E.3.3 compares it with alternative bag-level summaries.

Table 2: Component ablations on TCGA-KIRP and PANDA. Values are mean±std over five folds, and bold marks the highest mean.

	TCGA-KIRP		PANDA	
Model	**ACC**	**F1-Score**	**ACC**	**F1-Score**
QDMIL (full)	$\mathbf{0.8878 \pm 0.0389}$	$\mathbf{0.8302 \pm 0.0576}$	$\mathbf{0.7636 \pm 0.0058}$	$\mathbf{0.7244 \pm 0.0105}$
w/o conditioned update	0.8692 ± 0.0503	0.8107 ± 0.0716	0.7450 ± 0.0082	0.6968 ± 0.0135
w/o quantile-grid encoding	0.8765 ± 0.0360	0.8143 ± 0.0600	0.7316 ± 0.0026	0.6808 ± 0.0075
w/o quantile coordinates	0.8766 ± 0.0331	0.8163 ± 0.0480	0.7537 ± 0.0023	0.7146 ± 0.0057
w/o FiLM modulation	0.8692 ± 0.0449	0.8003 ± 0.0662	0.7455 ± 0.0038	0.7032 ± 0.0073

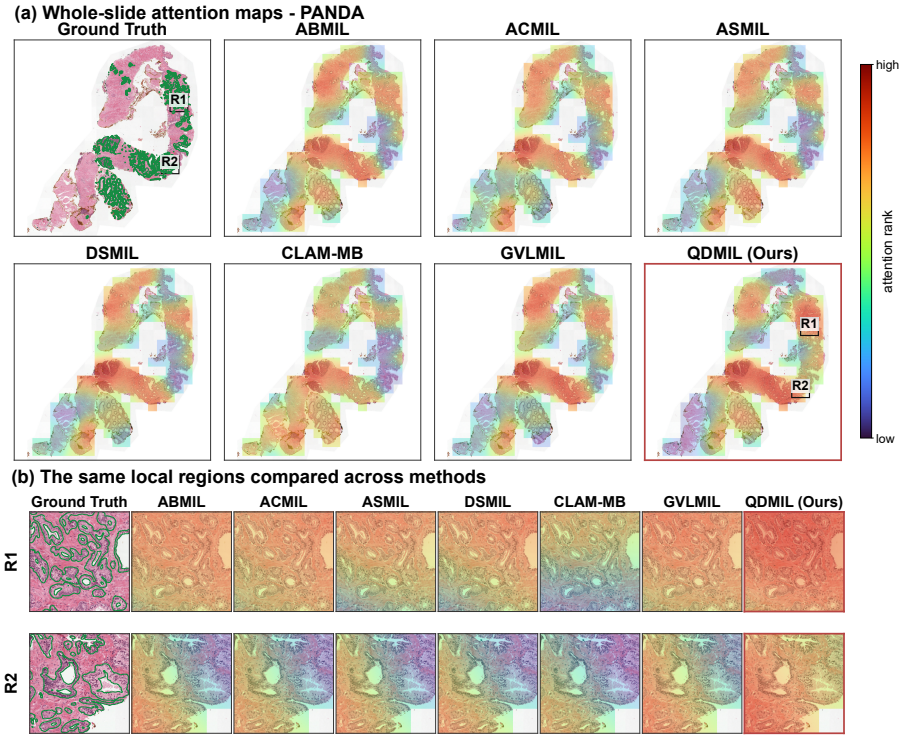

Figure 3: Whole-slide and local attention on PANDA. (a) Ground truth and seven model attention maps for one held-out slide. (b) Matched regions R1 and R2. Green contours mark cancerous epithelium (Gleason patterns 3–5). Warmer colors indicate higher within-slide attention percentiles.

4.4 Visualization

Figure 3 compares QDMIL with six baselines on a held out PANDA slide that all seven models classify correctly. The shared prediction shifts the comparison from outcome to regional prioritization. In the matched regions R1 and R2, QDMIL assigns greater relative attention to annotated cancerous

epithelium, while the baselines respond more weakly. For this slide, QDMIL's attention is therefore more closely aligned with the annotated tumor regions. We treat attention maps as qualitative evidence of model prioritization, not causal explanations. Additional examples on TCGA-KIRP, BRACS, and Brain-Mets are provided in Appendix E.2.

4.5 Parameter Sensitivity

K controls the number of learned projection directions, while M determines the resolution of each quantile summary. We vary them separately across the four datasets to test how much distributional resolution QDMIL requires. The reference value $K = 64$ attains or ties the highest mean ACC on all four datasets, while $M = 17$ does so on three (Figure 4a,b). Across all displayed settings, mean ACC remains within 2.69 percentage points of the reference. QDMIL is therefore not dependent on a narrow choice of K or M. Increasing either value beyond the reference offers no consistent benefit, indicating that a moderate grid is sufficient to represent task-relevant morphological variation. Appendix E.1 reports the complete sweeps, including the initial temperature T_0.

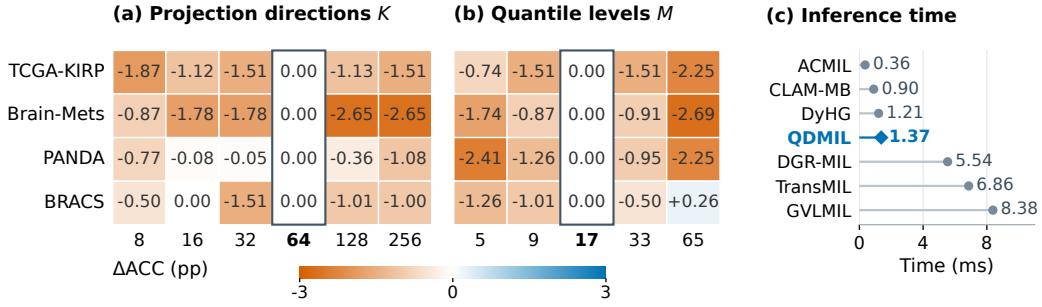

Figure 4: Parameter sensitivity and inference efficiency. (a,b) Changes in five-fold mean ACC relative to the $K = 64$ and $M = 17$ reference settings, with other parameters fixed. Outlined columns mark these settings. (c) Forward inference time on a synthetic bag of $N = 10000$ patches.

4.6 Computational Efficiency

Efficiency matters because a WSI can contain thousands of patches. Figure 4c benchmarks a forward pass on a synthetic bag with $N = 10000$ patches. QDMIL requires 1.37 ms per forward pass. The two attention-based baselines are faster, but neither matches its overall predictive performance in Table 1. QDMIL remains close to DyHG's 1.21 ms and runs approximately 4.0–$6.1\times$ faster than DGR-MIL, TransMIL, and GVLMIL. The fixed-size quantile grid thus provides slide-wide context without the heavier patch-indexed or patch-to-patch operators used by these alternatives. Appendix D reports the complete timing, memory, and FLOPs results, together with the measurement protocol and bag-size scaling analysis.

5 Conclusion

Our central conclusion is that permutation invariance in WSI is determined by how context is addressed rather than by patch interaction itself. Arbitrary patch indices can make an otherwise set-based pipeline order-sensitive, whereas the within-slide distribution provides a reference intrinsic to the bag. QDMIL applies this idea through a quantile grid that links local diagnostic evidence to slide-wide morphological composition. The theory provides permutation guarantees, while paired order controls and experiments across four pathology datasets demonstrate the practical value of this reference. Taken together, these findings support a broader design principle: context in an unordered slide should arise from slide content rather than patch enumeration. Building on this principle, future work can extend QDMIL in two directions. First, adaptive projection and quantile budgets could match distributional resolution to slide-level morphological complexity. Second, a complementary question is when spatial organization adds information beyond slide-wide morphological composition and how to incorporate it without index-based addressing. Both directions would extend the distributional reference while preserving its slide-intrinsic, permutation-invariant construction.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, support literature retrieval and discovery, and assist with research ideation and execution. The authors reviewed all resulting material and take full responsibility for the paper.

ETHICS STATEMENT

This study is a secondary analysis of four publicly released pathology datasets and did not recruit participants, collect new specimens, or involve interaction with patients. The original collection, consent, and ethical-governance procedures remain those specified by the respective dataset providers. We used the data solely for research benchmarking and did not attempt to identify any individual or release additional patient-level information. Differences in patient populations, acquisition sites, tissue preparation, annotation protocols, and class prevalence may introduce biases and limit generalization. The reported results do not establish clinical validity or equitable performance across demographic or institutional subgroups, and deployment would require external validation, privacy safeguards, and appropriate clinical oversight. We take responsibility for the ethical conduct and presentation of this work.

REPRODUCIBILITY STATEMENT

Section 3 describes the QDMIL architecture, learning objective, and forward computation, with interpolation details and complete proofs provided in Appendices B.1 and B.2. Dataset statistics, split construction, and training settings are documented in Appendix A, while the evaluation protocol is described in Section 4.1. The randomization and aggregation procedures for the order test are specified in Appendix C. Component ablations on TCGA-KIRP and PANDA are defined in Section 4.3. The parameter sweeps and the efficiency measurement procedure are provided in Appendices E.1 and D, respectively.

REFERENCES

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, et al. BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images. *Database*, 2022:baac093, 2022. doi: 10.1093/database/baac093.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Chen, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, 2022. doi: 10.1038/s41591-021-01620-2.

Linghan Cai, Shenjin Huang, Ye Zhang, et al. AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis*, 2025. doi: 10.1016/j.media.2025.103631.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck-Krauss-Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. *The New England Journal of Medicine*, 374(2):135–145, 2016. doi: 10.1056/NEJMoa1505917.

Saahil Chadha, Durga V. Sritharan, Darin Dolezal, et al. Matched MRI, segmentations, and histopathologic images of brain metastases from primary lung cancer. *Scientific Data*, 13:40, 2026. doi: 10.1038/s41597-025-06353-2.

Yuan-Chih Chen and Chun-Shien Lu. RankMix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuxuan Chen, Jiawen Li, Lianghui Zhu, et al. Dynamic hypergraph representation for bone metastasis analysis. *Computer Methods and Programs in Biomedicine*, 271:108966, 2025. doi: 10.1016/j.cmpb.2025.108966.

Philip Chikontwe, Meejeong Kim, Jaehoon Jeong, et al. FR-MIL: Distribution re-calibration-based multiple instance learning with transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 44(1):409–421, 2025. doi: 10.1109/TMI.2024.3446716.

Heather D. Couture, J. S. Marron, Charles M. Perou, Melissa A. Troester, and Marc Niethammer. Multiple instance learning for heterogeneous images: Training a CNN for histopathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11071 of *Lecture Notes in Computer Science*, pp. 254–262. Springer, 2018. doi: 10.1007/978-3-030-00934-2_29.

Heng Fang, Sheng Huang, Wenhao Tang, Luwen Huangfu, and Bo Liu. SAM-MIL: A spatial contextual aware multiple instance learning approach for whole slide image classification. In *ACM Multimedia*, 2024.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Juho Lee, Yoonho Lee, Jungtaek Kim, et al. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753, 2019.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2021. doi: 10.1109/CVPR46437.2021.01409.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, et al. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 2021. doi: 10.1038/s41551-020-00682-w.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Ethan Perez, Florian Strub, Harm de Vries, et al. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Daniel Shao, Joel Runevic, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, Andrew H. Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations*, 2026.

Zhuchen Shao, Hao Bian, Yang Chen, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos N. Plataniotis. ASMIL: Attention-stabilized multiple instance learning for whole slide imaging. In *International Conference on Learning Representations*, 2026.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, et al. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

Hongrun Zhang, Yanda Meng, Yitian Zhao, et al. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01824.

Xiao Zhang, Guangshuang Tan, Jie Hu, et al. Generative vision-language multiple instance learning for weakly supervised neonatal fundus screening and reporting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. FSPool: Learning set representations with featurewise sort pooling. In *International Conference on Learning Representations*, 2020.

Yunlong Zhang, Honglin Li, Yunxuan Sun, et al. Attention-challenging multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 125–143, 2024.

Wenhui Zhu, Xiwen Chen, Peijie Qiu, et al. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 333–351, 2024.

APPENDIX OVERVIEW

The appendix follows the logic of the main paper: it first specifies the data, then expands the algorithmic and theoretical foundations, tests the central permutation-invariance claim, documents reproducibility and measurement details, and finally provides supporting empirical analyses.

Contents.

- **Appendix A** documents the datasets, preprocessing, split construction, and training setup.
- **Appendix B** gives the quantile-interpolation details and complete proofs.
- **Appendix C** evaluates patch-order sensitivity and attributes the observed behavior to operator-level design.
- **Appendix D** derives the computational cost and documents the efficiency-measurement protocol.
- **Appendix E** collects parameter analyses, cross-dataset visualizations, and additional quantitative evidence.

A DATA AND EXPERIMENTAL SETUP

Cohorts and labels. Table 3 summarizes the four public WSI cohorts. TCGA-KIRP (Cancer Genome Atlas Research Network, 2016) groups AJCC pathologic stages I–II as early and III–IV as advanced. PANDA (Bulten et al., 2022) uses ISUP grades 0–5 from the public training set. BRACS (Brancati et al., 2022) uses benign, atypical, and malignant lesion groups. Brain-Mets (Chadha et al., 2026) separates small-cell lung carcinoma (SCLC), adenocarcinoma, and other lung-cancer histologies. The class counts in Table 3 follow these label orders.

Patch extraction and encoding. We extract non-overlapping patches at WSI level 0. Brain-Mets uses 256×256-pixel patches at $0.53\,\mu$m/px. The other datasets use 512×512-pixel patches, at $0.25\,\mu$m/px for TCGA-KIRP and BRACS and 0.45–$0.50\,\mu$m/px for PANDA. All patches are resized to 448×448 and encoded as 512-dimensional features by frozen CONCH. The resulting feature bags are shared across methods.

Fold construction. We construct five stratified folds from slide-level labels using `StratifiedKFold(n_splits=5, shuffle=False)` and reuse the assignments across methods and experimental variants. The 267 TCGA-KIRP slides represent 245 cases. After fold construction, we verify that no case appears in both the training and test partitions of any fold. Each Brain-Mets slide represents a distinct case. BRACS and PANDA remain split at the slide level because patient identifiers are unavailable in the releases used here.

Training setup. QDMIL uses a batch size of one slide and Adam (learning rate 2×10^{-4}, weight decay 10^{-5}), with 10 warmup epochs followed by cosine scheduling.

Table 3: Dataset composition. Class counts follow the label order defined above. Median bag sizes are estimated from 60 randomly sampled slides per dataset.

Dataset	Task	Classes	Slides	Class counts	Median bag size
TCGA-KIRP	Early/advanced stage	2	267	200/67	1.3×10^4
PANDA	ISUP grading	6	10611	2889/2664/1343/1242/1249/1224	1.3×10^2
BRACS	Breast lesion group	3	394	203/52/139	8.7×10^3
Brain-Mets	Lung-cancer histology	3	111	12/79/20	6.1×10^3

B ALGORITHMIC AND THEORETICAL DETAILS

B.1 QUANTILE INTERPOLATION

Let $N \geq 1$ and $M \geq 2$, and sort the projections along direction k as $s_{(1),k} \leq \cdots \leq s_{(N),k}$. For $m = 0, \ldots, M-1$, define $p_m = \tau_m(N-1)$, $j_m^- = \lfloor p_m \rfloor$, $j_m^+ = \lceil p_m \rceil$, and $\lambda_m = p_m - j_m^-$. The

interpolation in Eq. (3) is

$$Q_{k,m} = (1 - \lambda_m) s_{(j_m^- + 1),k} + \lambda_m s_{(j_m^+ + 1),k}. \tag{9}$$

When p_m is an integer, $j_m^- = j_m^+$ and $Q_{k,m}$ is the corresponding order statistic. When $N = 1$, every quantile equals the projection of the sole patch. The levels τ_m are fixed, and λ_m depends only on these levels and N. Neither contains learnable parameters. Sorting and interpolation are differentiable almost everywhere. Gradients flow directly through $s_{i,k}$ in Eq. (4) and through the order statistics contributing to Q. Tied projected scores form the usual nondifferentiable boundary of sorting, where the selected ordering can change.

B.2 Formal Statements and Proofs

Proof of Proposition 1. Order the computation graph topologically. The input patch sequence is permutation equivariant by definition. At each induction step, suppose that the available patch-level intermediates are equivariant and the available bag-level intermediates are invariant. A shared patch-wise mapping preserves equivariance. A symmetric reduction depends only on the multiset of its inputs, so applying it to an equivariant quantity produces an invariant quantity. Order statistics are invariant for the same reason. Combining an invariant bag-level quantity with each patch through a shared mapping again preserves equivariance. By induction, $\mathcal{G}$ is equivariant. The final prediction depends only on invariant bag-level quantities, so F is invariant. ∎

Example 1 (Distinguishability of internal distributional structure). Let $N = 4$ and consider two sets of projected scores along direction k, $A = \{0, 0, 1, 1\}$ and $B = \{0, \frac{1}{2}, \frac{1}{2}, 1\}$. Both have mean $\frac{1}{2}$, minimum 0, and maximum 1. For any $M \geq 4$, the first interior node satisfies $\tau_1 = 1/(M-1) \leq 1/3$. Linear interpolation gives $Q^A_{k,1} = 0$ and $Q^B_{k,1} = 3/[2(M-1)] > 0$. Thus, the quantile grids differ despite identical means and extrema.

In pathology, this distinction allows slides with the same average and extreme responses to retain differences in their internal morphological composition. The quantile grid therefore provides a finer bag-level reference for subsequent conditioning. Because WSI bag sizes vary substantially, the approximation error of this fixed-size reference should also remain controlled across N.

Proposition 2 (Approximation guarantee of the quantile grid). Let $N \geq 2$ and $M \geq 2$, and fix a direction k. The empirical quantile function satisfies $G_k(j/(N-1)) = s_{(j+1),k}$ for $j = 0, \ldots, N-1$, and we define the range along this direction as $R_k = s_{(N),k} - s_{(1),k}$. Let $G_{k,M}$ denote the piecewise-linear interpolation of G_k over the M uniformly spaced nodes in Eq. (3). If μ_k and $\mu_{k,M}$ have G_k and $G_{k,M}$ as their respective quantile functions, then

$$W_1(\mu_k, \mu_{k,M}) = \int_0^1 \left| G_k(t) - G_{k,M}(t) \right| \mathrm{d}t \ \leq \frac{R_k}{2(M-1)}. \tag{10}$$

Here, W_1 denotes the one-dimensional 1-Wasserstein distance. Proposition 2 bounds grid discretization after linear interpolation of the empirical quantile function. The next result accounts for the interpolation gap from the empirical measure itself.

Lemma 1 (Interpolation gap). Let $\hat{P}_N = \frac{1}{N} \sum_{j=1}^N \delta_{s_{(j),k}}$ be the empirical measure of the projected scores and $g_j = s_{(j+1),k} - s_{(j),k}$. Then

$$W_1(\hat{P}_N, \mu_k) = \sum_{j=1}^{N-1} g_j \, \frac{j^2 + (N-j)^2}{2N^2(N-1)} \ \leq \ \frac{R_k}{2N}, \tag{11}$$

with equality at $N = 2$. Combining with Proposition 2 through the triangle inequality,

$$W_1(\hat{P}_N, \mu_{k,M}) \ \leq \ \frac{R_k}{2N} + \frac{R_k}{2(M-1)}. \tag{12}$$

At the median bag sizes in Table 3 and the default $M = 17$, the grid term is about eight times larger than the interpolation term on PANDA and two to three orders of magnitude larger on the other datasets. Thus, grid resolution is the main contributor to this bound under the reported settings.

Corollary 1 (Approximation error of distributional functionals). If a real-valued functional Φ of a one-dimensional distribution is L-Lipschitz with respect to W_1, then

$$\left| \Phi(\hat{P}_N) - \Phi(\mu_{k,M}) \right| \leq L \left(\frac{R_k}{2N} + \frac{R_k}{2(M-1)} \right). \tag{13}$$

The fixed-size grid therefore controls the error of any such functional relative to the empirical projected-score distribution.

Proof of Proposition 2. G_k is nondecreasing, and $G_{k,M}$ agrees with it at every grid node and is linear between adjacent nodes. Let $\delta = 1/(M-1)$ and define $\Delta_m = G_k(\tau_{m+1}) - G_k(\tau_m) \geq 0$ for $m = 0, \ldots, M-2$. On $[\tau_m, \tau_{m+1}]$, the two functions share the same endpoints. Among monotone functions with these endpoints, a threshold step function maximizes the L_1 distance from the connecting line segment. If its step occurs at distance $c \in [0, \delta]$ from the left endpoint, this distance is $\Delta_m[c^2 + (\delta - c)^2]/(2\delta) \leq \delta\Delta_m/2$. Hence, $\int_{\tau_m}^{\tau_{m+1}} |G_k - G_{k,M}| \leq \delta\Delta_m/2$. Summing over intervals and using $\sum_m \Delta_m = G_k(1) - G_k(0) = R_k$ gives $\int_0^1 |G_k - G_{k,M}| \leq R_k/[2(M-1)]$. In one dimension, W_1 equals the integral of the absolute difference between the corresponding quantile functions, which proves the result. ■

Proof of Lemma 1. Because G_k is linear on $[\,(j-1)/(N-1),\, j/(N-1)\,]$, μ_k assigns mass $1/(N-1)$ uniformly to $[s_{(j),k}, s_{(j+1),k}]$. Write $x = s_{(j),k} + g_j\theta$ with $\theta \in [0,1]$. On this interval, the cumulative distribution functions of $\hat{P}_N$ and μ_k are j/N and $(j-1+\theta)/(N-1)$. Their signed difference is $[N(1-\theta) - j]/[N(N-1)]$, whose zero lies in $(0,1)$. Integrating its absolute value over θ and multiplying by g_j gives $g_j[\,j^2 + (N-j)^2\,]/[2N^2(N-1)]$. Summing over j and using the cumulative representation of one-dimensional W_1 proves the equality in Eq. (11). The coefficient is largest at $j = 1$ and $j = N-1$. Since $[1 + (N-1)^2]/[N(N-1)] \leq 1$ for $N \geq 2$ and $\sum_j g_j = R_k$, the upper bound is $R_k/(2N)$, with equality at $N = 2$. ■

Proof of Corollary 1. The Lipschitz condition and Eq. (12) give $|\Phi(\hat{P}_N) - \Phi(\mu_{k,M})| \leq LW_1(\hat{P}_N, \mu_{k,M}) \leq L\{R_k/(2N) + R_k/[2(M-1)]\}$. ■

Proof of Corollary 2. Under a patch permutation, the rows of E and S permute, while quantile extraction leaves Q unchanged. Each r_i is computed from its own projected scores and the invariant grid through the same mapping, so the sequences of coordinates and conditioned representations remain equivariant. The attention scores and weights permute with these representations, making their normalized weighted sum u invariant. The grid encoding q is also invariant because it depends only on Q. Thus, the conditioned patch sequence is equivariant and the fused logits from $[u; q]$ are invariant, as required by Proposition 1. ■

C Permutation-Invariance Evaluation

Proposition 1 concerns exact model behavior, so we test its practical counterpart by reordering patches at inference. Original-order repeat controls separate changes caused by permutation from ordinary inference-time variation.

We compare the original raster storage order with five fixed random permutations of each bag. ΔACC and ΔF1-Score measure changes from the original order. Relative logit change is first maximized across output components for each bag, then summarized across bags, permutations, and folds.

Table 4: Patch-order sensitivity on TCGA-KIRP and PANDA. Repeat is the median change between original-order reevaluations.

		TCGA-KIRP			PANDA		
Method	**Type**	**ΔACC**	**Median rel. $\lvert\Delta$logit$\rvert$**	**Repeat**	**ΔACC**	**Median rel. $\lvert\Delta$logit$\rvert$**	**Repeat**
QDMIL (ours)	**Bag-conditioned**	**+0.0000**	$\mathbf{8.3 \times 10^{-8}}$	**0**	**+0.0000**	$\mathbf{8.3 \times 10^{-8}}$	**0**
ACMIL	Self-scoring	+0.0000	8.7×10^{-8}	0	+0.0000	6.9×10^{-8}	0
ABMIL	Self-scoring	+0.0000	1.1×10^{-7}	0	+0.0000	7.9×10^{-8}	0
GVLMIL	Patch interaction	+0.0000	1.1×10^{-7}	0	+0.0000	7.9×10^{-8}	0
DyHG	Stochastic inference	+0.0000	2.6×10^{-4}	2.6×10^{-4}	−0.0009	2.3×10^{-2}	2.4×10^{-2}
DGR-MIL	Patch interaction	−0.0217	1.5×10^{-1}	0	−0.0058	8.3×10^{-2}	0
TransMIL	Patch interaction	−0.0354	3.8×10^{-1}	0	−0.0313	4.3×10^{-1}	0

The F1-Score results follow the same pattern. QDMIL, ACMIL, ABMIL, and GVLMIL remain unchanged on all four datasets. On BRACS, where class imbalance makes F1-Score particularly in-

formative, the changes for DyHG, DGR-MIL, and TransMIL are -0.0007, -0.0270, and -0.0584, respectively.

Order-invariant interactions. QDMIL preserves both bag-level metrics under reordering across all four datasets. GVLMIL does so as well despite modeling full patch-to-patch interactions. Its median relative logit changes remain at the single-precision rounding scale observed for QDMIL, ABMIL, and ACMIL. The bag mean, pairwise score matrix, value-selected top-k neighborhoods, and symmetric aggregation in GVLMIL are invariant or equivariant. Indices enter only to break ties during top-k selection. This comparison shows that patch interactions alone do not cause order sensitivity.

Index-dependent operators. TransMIL and DGR-MIL contain operators that address the patch sequence by index. TransMIL's PPEG reshapes the sequence into a pseudo-spatial grid before applying depthwise convolution, while DGR-MIL performs linear compression along the patch axis. Their ACC decreases by 3.54 and 2.17 percentage points on TCGA-KIRP, respectively. TransMIL also loses 3.13 points on PANDA, and both methods decline on Brain-Mets and BRACS. The pattern therefore spans cohorts with markedly different bag sizes and matches the addressing principle in Proposition 1.

Repeat controls. Repeat controls use two consecutive passes without reordering and the same aggregation. All methods except DyHG produce identical logits across these passes. For QDMIL, ACMIL, ABMIL, and GVLMIL, the remaining median changes under permutation are approximately 10^{-7} or smaller and are consistent with numerical rounding. DyHG instead constructs a dynamic hypergraph with Gumbel-Softmax, which samples noise during inference. On BRACS, its repeat control yields a median relative logit change of 8.8×10^{-4} and a maximum of 0.19, compared with 8.1×10^{-4} and 0.18 under permutation. The permutation test alone therefore cannot distinguish an order effect from inference-time variation for DyHG. This is why the repeat control is needed.

Table 5: Patch-order sensitivity on Brain-Mets and BRACS.

		Brain-Mets			BRACS		
Method	**Type**	**ΔACC**	**Median rel. |Δlogit|**	**Repeat**	**ΔACC**	**Median rel. |Δlogit|**	**Repeat**
QDMIL (ours)	**Bag-conditioned**	**+0.0000**	$\mathbf{6.3 \times 10^{-8}}$	**0**	**+0.0000**	$\mathbf{7.6 \times 10^{-8}}$	**0**
ACMIL	Self-scoring	+0.0000	8.0×10^{-8}	0	+0.0000	8.0×10^{-8}	0
ABMIL	Self-scoring	+0.0000	9.6×10^{-8}	0	+0.0000	1.0×10^{-7}	0
GVLMIL	Patch interaction	+0.0000	8.7×10^{-8}	0	+0.0000	8.9×10^{-8}	0
DyHG	Stochastic inference	+0.0000	4.4×10^{-4}	4.4×10^{-4}	−0.0010	8.1×10^{-4}	8.8×10^{-4}
DGR-MIL	Patch interaction	−0.0218	3.3×10^{-2}	0	−0.0268	7.3×10^{-2}	0
TransMIL	Patch interaction	−0.0164	1.8×10^{-1}	0	−0.0223	2.1×10^{-1}	0

D Computational Cost and Efficiency Measurement

Per-module cost. The input mapping and directional projection cost $O(Ndh)$ and $O(NhK)$, respectively. Quantile sorting costs $O(NK \log N)$, while soft assignment costs $O(NKM)$. Coordinate encoding and attention scoring require $O(N(Kh' + h'h))$ and $O(Nha)$, respectively. The bag-level affine mapping and quantile-grid encoding operate on the fixed-size grid. Their costs depend on K, M, and the model widths but not on N. Thus, with d, K, M, h, h', a, and C fixed, the forward-pass complexity is $O(N \log N)$. No term is quadratic in the number of patches.

Measurement protocol. We report peak allocated GPU memory in MiB. Runtime is averaged over five forward passes after two warmup passes. Table 7 reports the mean of two independent runs, and the parameter and FLOP counts correspond to models with three output classes. QDMIL uses $K = 64$ and $M = 17$ throughout these measurements.

Memory scaling. The largest intermediate introduced by the quantile-coordinate computation is the $N \times K \times M$ soft-assignment tensor. Its storage is linear in N when K and M are fixed, although the measured peak also includes parameters and the remaining intermediate tensors. Figure 5 and Table 6 are consistent with this scaling over bags from 1000 to 32000 patches. QDMIL's memory increases by $1.86\times$ from 16000 to 32000 patches, close to the factor expected under linear growth.

Its peak memory remains below that of TransMIL and GVLMIL throughout the tested range. The additional soft-assignment tensor contributes to its higher footprint than ABMIL and DyHG. **Full comparison.** Table 7 extends the benchmark to all 15 baselines under the same setup.

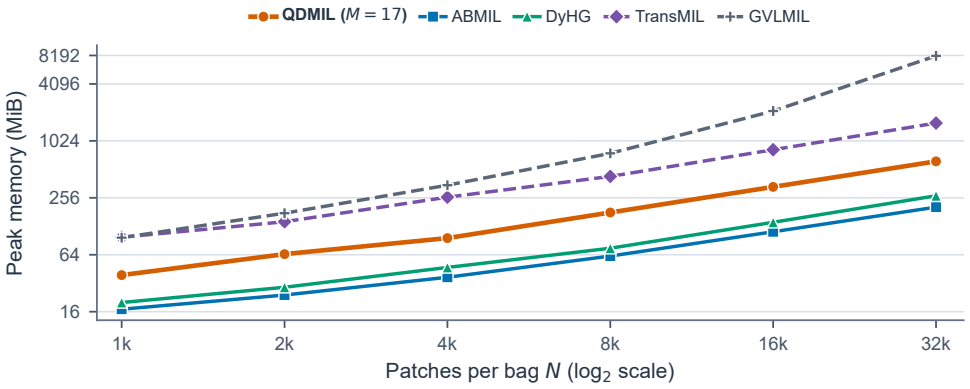

Figure 5: Peak inference memory versus bag size on base-2 logarithmic axes.

Table 6: Peak inference memory (MiB) across bag sizes. The final column gives the 16k-to-32k growth factor.

Method	1k	2k	4k	8k	16k	32k	16k→32k
QDMIL (ours)	**39**	**65**	**96**	**179**	**334**	**622**	**1.86×**
ABMIL	17	24	37	62	112	204	1.83×
DyHG	20	29	47	75	141	269	1.90×
TransMIL	98	143	260	431	825	1580	1.92×
GVLMIL	97	176	348	756	2118	8108	3.83×

Table 7: Inference efficiency for a synthetic bag with $N = 10{,}000$, 512-dimensional features, and three classes. Methods are ordered by inference time.

Method	Params. (M)	Inference Time (ms)	Peak Memory (MiB)	FLOPs (G)
ACMIL	0.26	0.36	55	1.97
ABMIL	0.33	0.46	71	3.28
IBMIL	0.41	0.46	100	3.28
RankMix	0.33	0.52	93	3.28
CLAM-SB	0.53	0.78	91	5.25
CLAM-MB	0.53	0.90	90	5.25
DyHG	0.34	1.21	107	3.34
DSMIL	0.09	1.35	56	0.83
QDMIL (ours)	**1.68**	**1.37**	**210**	**5.43**
ASMIL	1.72	1.47	74	7.88
ASMIL-EMA	1.72	1.52	74	7.88
FR-MIL	1.32	1.62	216	7.92
DTFD-MIL	0.53	2.75	94	5.25
DGR-MIL	4.08	5.54	525	21.73
TransMIL	2.41	6.86	529	24.93
GVLMIL	1.84	8.38	1028	20.97

E SUPPLEMENTARY EMPIRICAL ANALYSES

E.1 PARAMETER SENSITIVITY

Figure 6 examines the direction count K, the number of quantile levels M, and the initial temperature T_0. Across all sweep points, the ranges of mean ACC on TCGA-KIRP, Brain-Mets, PANDA, and BRACS are 0.0299, 0.0269, 0.0280, and 0.0177, respectively. Performance therefore remains within three percentage points across each full set of tested configurations despite changes in grid width, grid resolution, and temperature initialization.

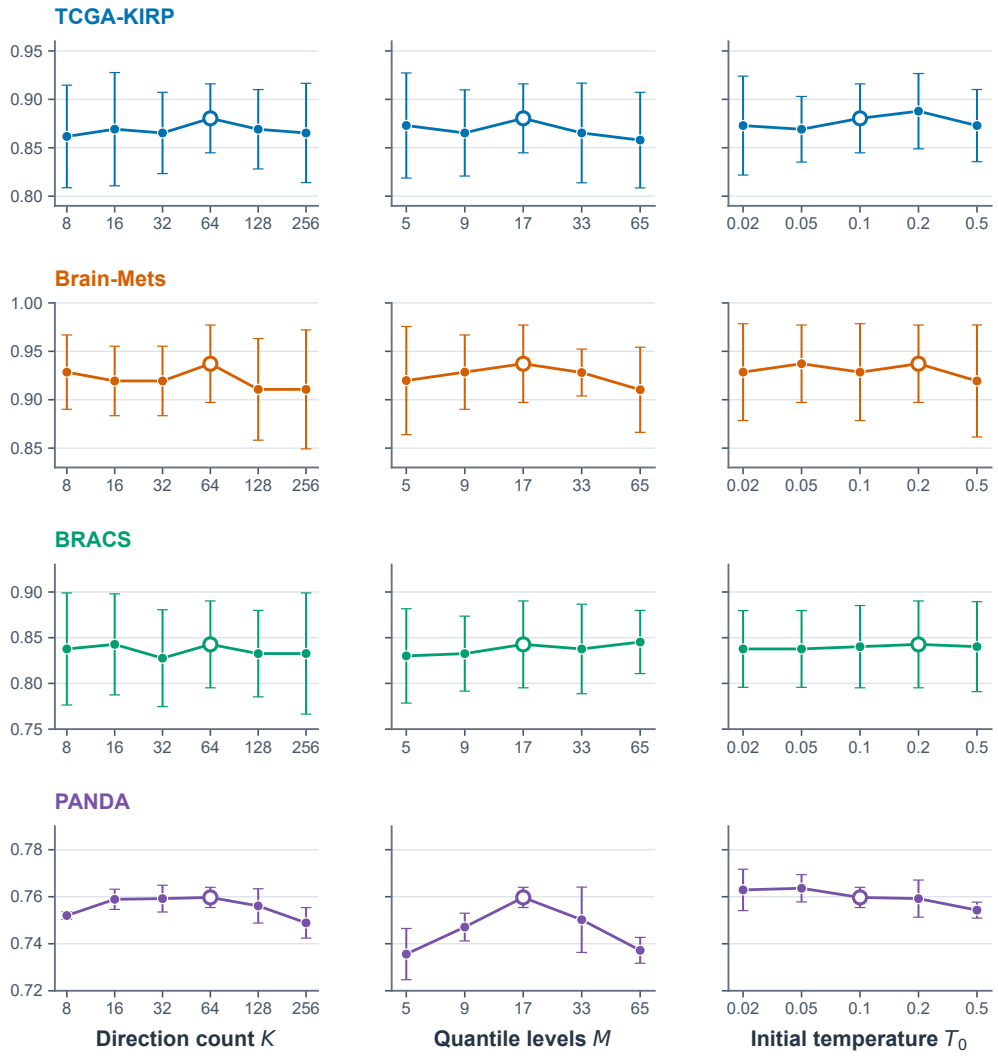

Figure 6: One-parameter sensitivity of ACC to K, M, and T_0. Points and error bars show five-fold mean±sample standard deviation, and open markers denote the reference configurations. Lines connect discrete settings only as visual guides. Rows use separate vertical ranges.

$K = 64$ attains the highest ACC on three datasets and ties for the highest on BRACS. Likewise, $M = 17$ is best on three datasets and lies within 0.0026 of the best result on BRACS. Increasing either parameter further provides no consistent gain, so the shared $K = 64$, $M = 17$ reference offers sufficient resolution without needlessly enlarging the grid. Consistent with learning the temperature after initialization, T_0 has the smallest or tied-smallest ACC range among the three parameters on every dataset. On PANDA, for example, varying T_0 from 0.02 to 0.5 changes ACC by only 0.0093.

The bound in Appendix B.2 measures how faithfully the grid approximates the bag distribution, not how downstream accuracy changes with resolution. The non-monotonic response to M is therefore not a contradiction. It shows that finer distributional approximation beyond a moderate resolution does not necessarily improve classification.

Table 8 reports ACC and F1-Score for every tested setting as five-fold mean±sample standard deviation. Each experiment varies one parameter while holding the other two at the reference configuration. The reference uses $K = 64$ and $M = 17$, with $T_0 = 0.1$ for TCGA-KIRP and PANDA and $T_0 = 0.2$ for BRACS and Brain-Mets. Bold means marked by † identify these reference points.

Table 8: Complete one-parameter sweeps on four datasets.

		TCGA-KIRP		**PANDA**		**BRACS**		**Brain-Mets**	
Parameter	**Value**	**ACC**	**F1-Score**	**ACC**	**F1-Score**	**ACC**	**F1-Score**	**ACC**	**F1-Score**
K	8	0.8617 ±0.0530	0.7895 ±0.0961	0.7520 ±0.0016	0.7107 ±0.0039	0.8377 ±0.0613	0.7575 ±0.1017	0.9285 ±0.0384	0.8951 ±0.0462
	16	0.8692 ±0.0585	0.8201 ±0.0762	0.7589 ±0.0043	0.7209 ±0.0062	0.8427 ±0.0553	0.7494 ±0.0921	0.9194 ±0.0359	0.8776 ±0.0517
	32	0.8653 ±0.0420	0.8020 ±0.0593	0.7592 ±0.0057	0.7206 ±0.0096	0.8276 ±0.0530	0.7243 ±0.0779	0.9194 ±0.0359	0.8754 ±0.0467
	64	**0.8804**† ±0.0356	**0.8211**† ±0.0505	**0.7597**† ±0.0043	**0.7207**† ±0.0077	**0.8427**† ±0.0475	**0.7674**† ±0.0597	**0.9372**† ±0.0400	**0.9147**† ±0.0561
	128	0.8691 ±0.0410	0.8133 ±0.0569	0.7561 ±0.0073	0.7139 ±0.0099	0.8326 ±0.0473	0.7240 ±0.0613	0.9107 ±0.0525	0.8623 ±0.0717
	256	0.8653 ±0.0513	0.8059 ±0.0726	0.7489 ±0.0065	0.7041 ±0.0118	0.8327 ±0.0663	0.7499 ±0.0791	0.9107 ±0.0615	0.8565 ±0.0947
M	5	0.8730 ±0.0543	0.8205 ±0.0696	0.7356 ±0.0109	0.6831 ±0.0229	0.8301 ±0.0516	0.7364 ±0.0678	0.9198 ±0.0559	0.8867 ±0.0713
	9	0.8653 ±0.0445	0.7876 ±0.0967	0.7471 ±0.0059	0.7026 ±0.0092	0.8326 ±0.0410	0.7527 ±0.0535	0.9285 ±0.0384	0.8951 ±0.0462
	17	**0.8804**† ±0.0356	**0.8211**† ±0.0505	**0.7597**† ±0.0043	**0.7207**† ±0.0077	**0.8427**† ±0.0475	**0.7674**† ±0.0597	**0.9372**† ±0.0400	**0.9147**† ±0.0561
	33	0.8653 ±0.0515	0.7997 ±0.0936	0.7502 ±0.0139	0.7073 ±0.0208	0.8377 ±0.0490	0.7554 ±0.0827	0.9281 ±0.0242	0.8918 ±0.0297
	65	0.8579 ±0.0494	0.7906 ±0.0656	0.7372 ±0.0055	0.6840 ±0.0065	0.8453 ±0.0346	0.7566 ±0.0720	0.9103 ±0.0440	0.8639 ±0.0618
T_0	0.02	0.8729 ±0.0511	0.8080 ±0.0795	0.7629 ±0.0088	0.7217 ±0.0115	0.8377 ±0.0420	0.7457 ±0.0512	0.9285 ±0.0500	0.9034 ±0.0663
	0.05	0.8691 ±0.0339	0.8082 ±0.0391	0.7636 ±0.0058	0.7244 ±0.0105	0.8377 ±0.0420	0.7581 ±0.0533	0.9372 ±0.0400	0.9147 ±0.0561
	0.1	**0.8804**† ±0.0356	**0.8211**† ±0.0505	**0.7597**† ±0.0043	**0.7207**† ±0.0077	0.8402 ±0.0450	0.7585 ±0.0586	0.9285 ±0.0500	0.9034 ±0.0663
	0.2	0.8878 ±0.0389	0.8302 ±0.0576	0.7592 ±0.0079	0.7197 ±0.0131	**0.8427**† ±0.0475	**0.7674**† ±0.0597	**0.9372**† ±0.0400	**0.9147**† ±0.0561
	0.5	0.8729 ±0.0373	0.8149 ±0.0456	0.7543 ±0.0034	0.7126 ±0.0058	0.8402 ±0.0492	0.7581 ±0.0567	0.9194 ±0.0579	0.8838 ±0.0893

E.2 Cross-Dataset Attention Visualizations

We extend the visualization in Section 4.4 to the other three datasets using the configurations from the main comparison. The held-out examples have compact tissue regions and were selected to expose cases in which the models reach different slide-level decisions. All maps use the same within-slide percentile normalization, spatial smoothing, opacity, and matched local windows. The local crops retain the whole-slide normalization. For DSMIL and CLAM-MB, we display the attention branch for the predicted class. Because QDMIL also uses a separate distributional readout, its map represents the local evidence pathway rather than the entire prediction mechanism.

E.2.1 TCGA-KIRP

Figure 7 shows an early-stage slide correctly classified by QDMIL, whereas all six displayed baselines predict advanced stage. Several methods still emphasize visually salient tissue. This contrast separates local saliency from slide-level interpretation: identifying prominent regions is insufficient if their significance is not assessed relative to the slide-wide morphology. Regional labels are unavailable, so this comparison is qualitative rather than a localization evaluation.

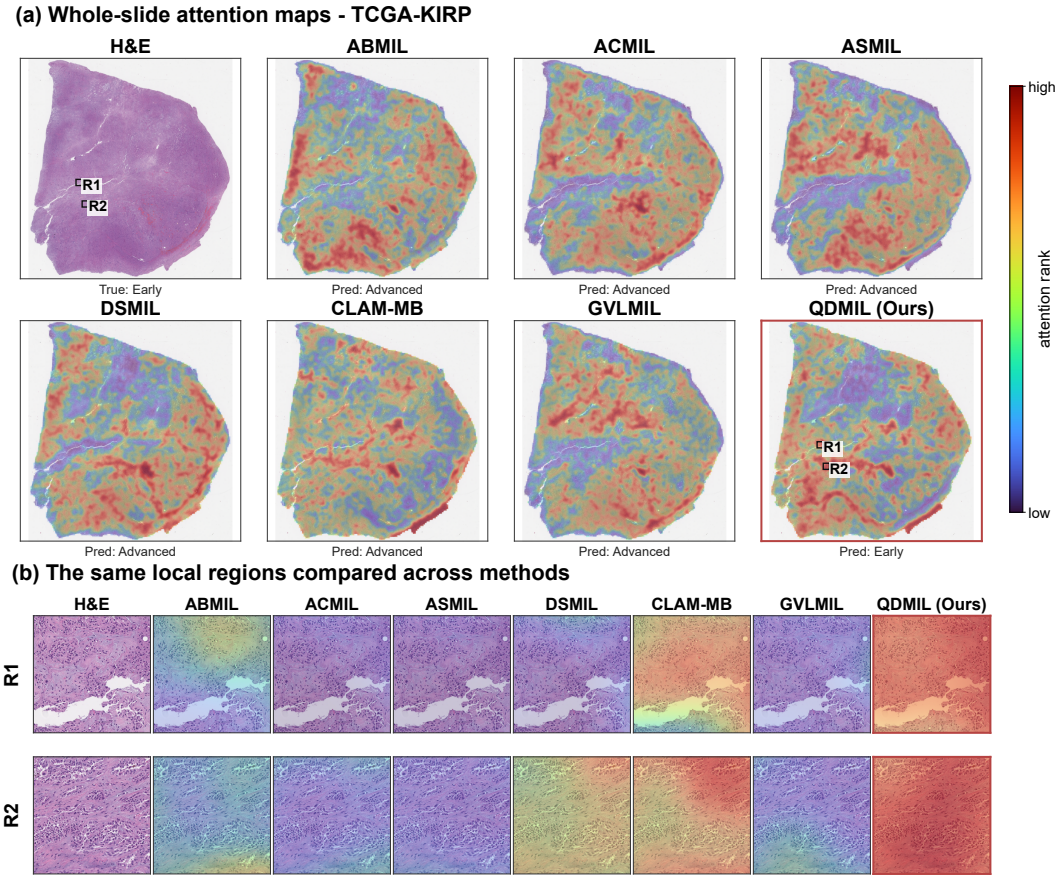

Figure 7: Attention maps for a held-out TCGA-KIRP slide (patient TCGA-P4-AAVO, fold 4). (a) H&E reference and model maps with slide labels and predictions. (b) Matched 2048×2048-pixel crops. Regional annotations are unavailable.

E.2.2 BRACS

Figure 8 shows an atypical slide correctly classified by QDMIL, while all six displayed baselines predict benign. The released contours identify atypical ductal hyperplasia (ADH) and flat epithelial atypia (FEA). Several methods attend to the same structures but reach different decisions. The example shows that agreement in local saliency does not guarantee the same bag-level interpretation. QDMIL's prediction is consistent with using the separate distributional readout to place this local evidence in slide-wide context.

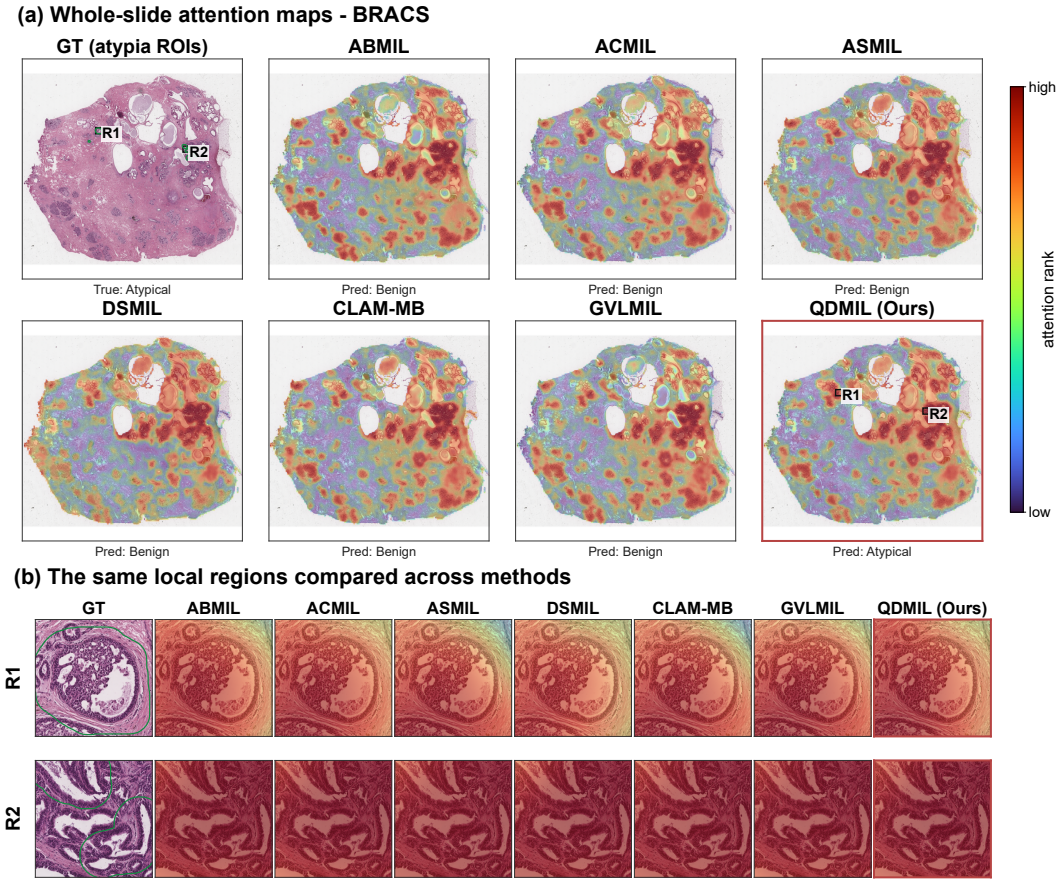

Figure 8: Attention maps for a held-out BRACS slide (BRACS_1899, fold 4). (a) H&E reference with released ADH/FEA contours and model maps. (b) Matched 2048×2048-pixel crops. The contours are selected annotations rather than exhaustive segmentation.

E.2.3 BRAIN-METS

Figure 9 shows a large-cell neuroendocrine carcinoma (LCNEC) example in the Other category of the three-class task. QDMIL and ACMIL predict this category correctly, whereas the other five displayed baselines predict small-cell lung carcinoma (SCLC). Because the methods inspect the same tissue and ACMIL also succeeds, the example does not attribute the result to a uniquely localized region. Instead, it extends the distinction between local evidence and bag-level interpretation to a heterogeneous residual category.

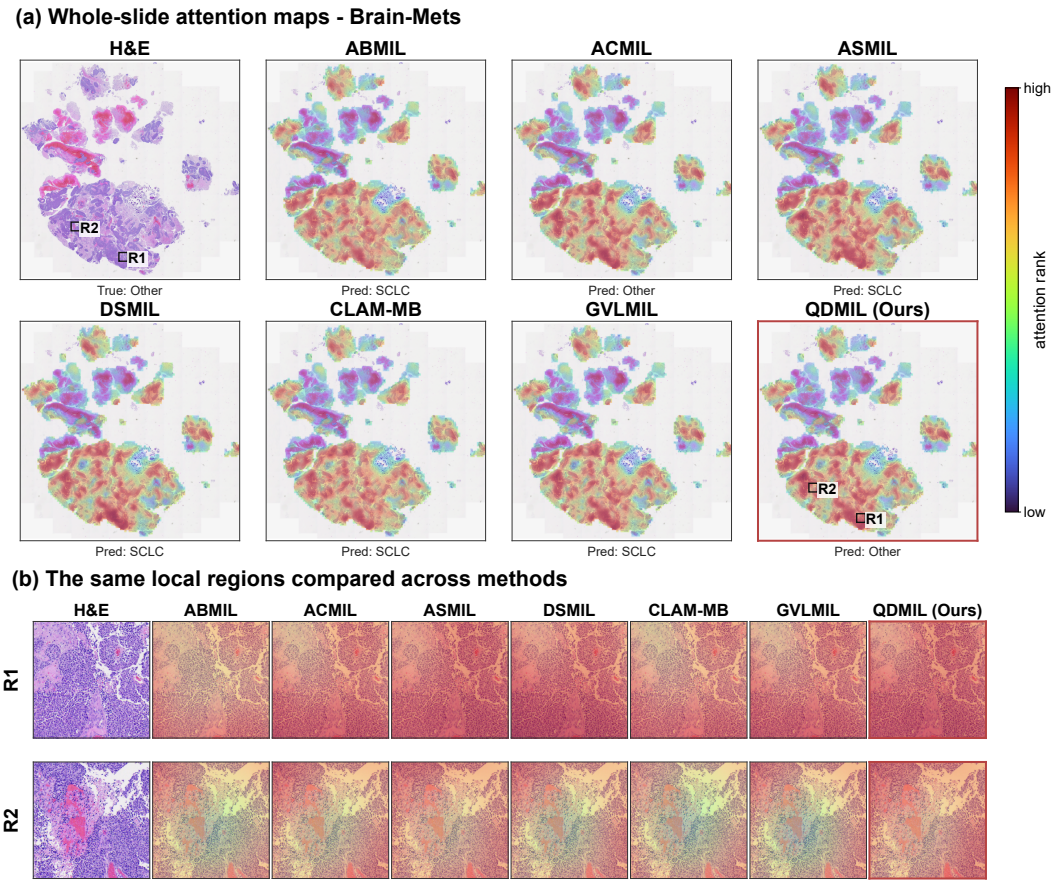

Figure 9: Attention maps for a held-out Brain-Mets slide (YG_9P6T37XA3XDG, fold 3). (a) H&E reference and model maps. (b) Matched 1024×1024-pixel crops. The clinical label is LCNEC within Other, and regional annotations are unavailable.

E.3 Additional Quantitative Analyses

The analyses below test whether the main results persist across classes, folds, bag sizes, and ordinal error severity. We also isolate the value of the quantile grid by replacing it with simpler bag summaries while preserving the downstream architecture.

E.3.1 Class-wise Analysis on PANDA

We compare QDMIL with DyHG, the strongest PANDA baseline in Table 1, using the saved held-out predictions from the same main-table runs. Pooling the five disjoint folds yields 10,611 slides, with 2,889, 2,664, 1,343, 1,242, 1,249, and 1,224 slides in ISUP grades 0–5, respectively. Figure 10 reports row-normalized confusion matrices and class-wise F1-Score computed from these pooled predictions.

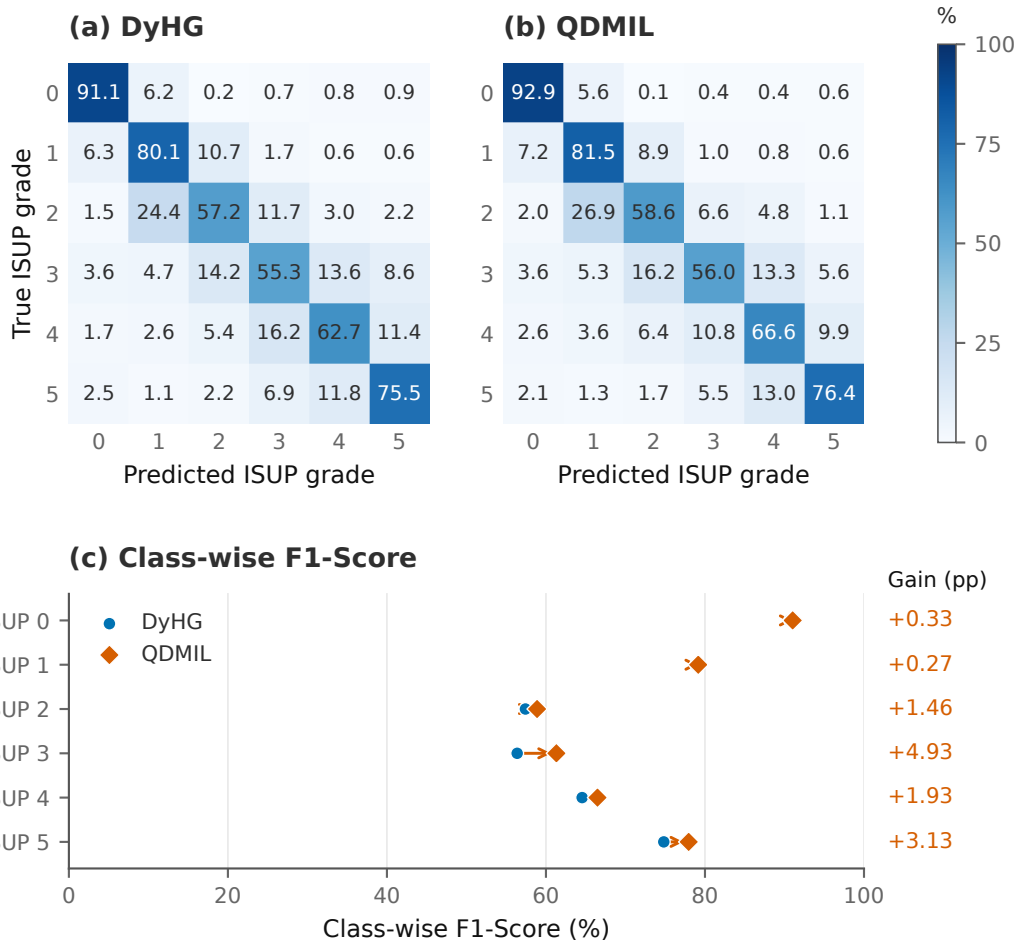

Figure 10: Class-wise PANDA results from pooled held-out predictions. (a,b) Row-normalized confusion matrices share one scale. (c) Per-class F1-Score and QDMIL–DyHG gains.

QDMIL improves F1-Score in all six grades, with the largest gain for ISUP 3, from 56.38% to 61.31% (+4.93 percentage points). The gains for ISUP 4 and 5 are 1.93 and 3.13 points, respectively. These improvements are larger than those for the two most prevalent grades, while diagonal recall increases in every grade. Thus, the aggregate gain is not driven only by the majority categories.

E.3.2 Paired Fold-wise Comparison on PANDA

Figure 11 shows all five paired differences from the main-table runs against DyHG, ABMIL, and CLAM-MB. QDMIL improves both metrics in every fold against ABMIL and CLAM-MB. Against DyHG, ACC improves in all five folds and F1-Score in four, with the remaining difference only -0.07 percentage points. This consistency shows that the overall advantage is not produced by a single favorable split.

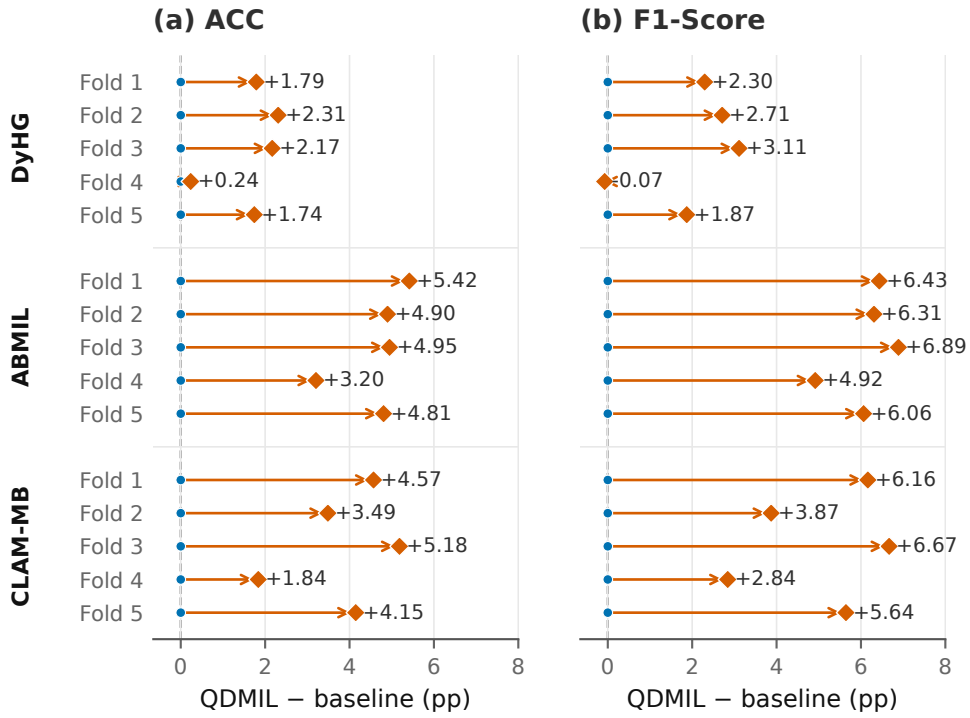

Figure 11: Paired QDMIL gains on five PANDA folds. Each marker is QDMIL minus the baseline in percentage points, with zero shown in blue.

E.3.3 Alternative Bag-level Summaries on TCGA-KIRP

We replace the quantile grid with either the mean or a triplet of mean, maximum, and standard deviation of the projected patch scores in each direction. The mean is broadcast, or the triplet is cyclically repeated, along the M slots to preserve the $K \times M$ input shape and network parameter count. Both replacements feed the conditioned update and the bag-level encoding branch. All other modules and the comparison settings ($K = 64$, $M = 17$, $T_0 = 0.2$) are retained. Table 9 reports mean$\pm$sample standard deviation over five folds.

Table 9: Alternative bag-level summaries on TCGA-KIRP.

Summary	**ACC**	**F1-Score**
Mean	0.8654 ± 0.0415	0.8005 ± 0.0436
Mean + max + SD	0.8691 ± 0.0410	0.8065 ± 0.0585
Quantile grid (QDMIL)	$\mathbf{0.8878 \pm 0.0389}$	$\mathbf{0.8302 \pm 0.0576}$

The quantile grid yields the highest mean ACC and F1-Score, exceeding the two replacements by 1.87–2.24 and 2.37–2.98 percentage points, respectively. Because the input shape, downstream modules, and parameter count are held fixed, the difference isolates the information retained by the

bag summary. The gain over mean, maximum, and standard deviation indicates that the internal organization of the patch distribution contributes beyond its central tendency, spread, and extrema.

E.3.4 Performance across Bag Sizes on PANDA

We group all PANDA slides by patch-count quartiles, keeping ties together, and recompute ACC and F1-Score from the pooled held-out predictions of the main-table runs. QDMIL improves both metrics over all three comparators in every quartile (Figure 12). The gains are largest in the smallest-bag group, spanning 6–109 patches, for both metrics against every baseline. The benefit is not confined to slides with many patches. Because the quartiles can also differ in class composition, this analysis establishes consistency across observed bag sizes rather than a causal effect of bag size.

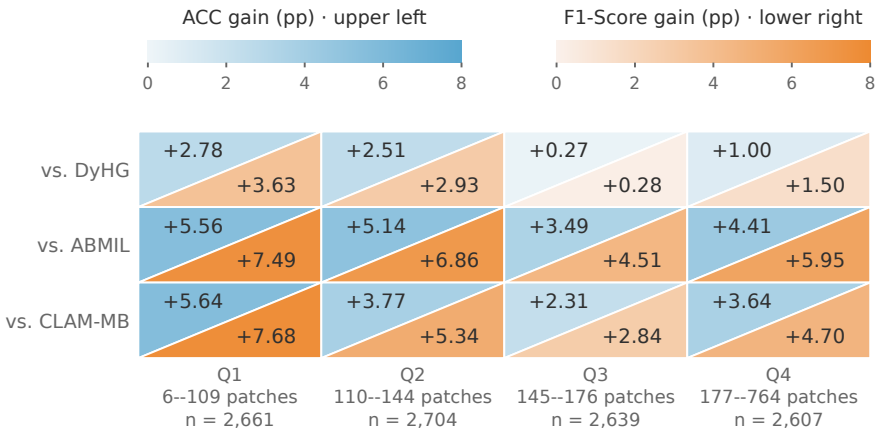

Figure 12: QDMIL gains across PANDA bag-size quartiles. Cells report ACC (upper left) and F1-Score (lower right) differences from each baseline in percentage points.

E.3.5 Ordinal Grading Errors on PANDA

Because PANDA grades are ordinal, we also examine the magnitude of misclassification. QDMIL has the fewest one-grade and two-or-more-grade errors among the compared methods (Figure 13), together with the highest quadratic-weighted kappa. Relative to DyHG, the more severe errors decrease from 731 to 682. Thus, the higher exact accuracy is not obtained by shifting mistakes farther across the grading scale.

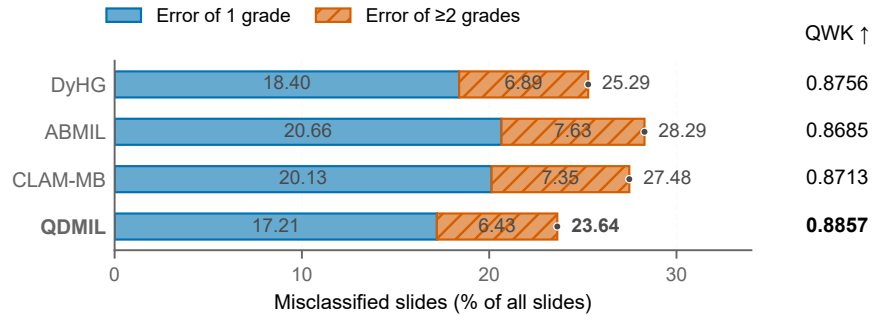

Figure 13: Ordinal errors on PANDA. Stacks show percentages of all 10,611 slides, and endpoint labels give total error rates. QWK is computed from the pooled held-out predictions.